Amendments to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (currently amended): A system for grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:

a scoring module determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is based on at least one of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight; [[and]]

a clustering module forming clusters of the documents by ~~applying~~ evaluating the score for the at least one concept [[to]] of each document for a best fit ~~criterion for each such document.~~ to the clusters and assigning each document to the cluster with the best fit; and

a threshold module dynamically determining a threshold for each cluster based on similarities between the documents grouped into the cluster and a center of the cluster, and reassigning those documents having similarities outside the threshold.

2. (original): A system according to Claim 1, further comprising:

the scoring module calculating the score as a function of a summation of at least one of the frequency of occurrence, the concept weight, the structural weight, and the corpus weight of the at least one concept.

3. (original): A system according to Claim 2, further comprising:

a compression module compressing the score through logarithmic compression.

4. (original): A system according to Claim 1, further comprising:

the scoring module calculating the concept weight as a function of a number of terms comprising the at least one concept.

5. (original): A system according to Claim 1, further comprising:
the scoring module calculating the structural weight as a function of a location of the at least one concept within the at least one such document.

6. (original): A system according to Claim 1, further comprising:
the scoring module calculating the corpus weight as a function of a reference count of the at least one concept over the plurality of documents.

7. (original): A system according to Claim 1, further comprising:
the scoring module forming the score assigned to the at least one concept to a normalized score vector for each such document, determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector, and applying the similarity to the best fit criterion.

8. (original): A system according to Claim 1, further comprising:
the clustering module evaluating a set of candidate seed documents selected from the plurality of documents, identifying a set of seed documents by applying the score for the at least one concept to a best fit criterion for each such candidate seed document, and basing the best fit criterion on the score of each such seed document.

9. (currently amended): A method for grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:
determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is based on at least one of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight; [[and]]

forming logically-grouped clusters of the documents by ~~applying~~ evaluating the score for the at least one concept [[to]] of each document for a best fit ~~criterion for each such document.~~ to the clusters and assigning each document to the cluster with the best fit;

dynamically determining a threshold for each cluster based on similarities between the documents grouped into the cluster and a center of the cluster; and

reassigning those documents having similarities outside the threshold.

10. (original): A method according to Claim 9, further comprising:

calculating the score as a function of a summation of at least one of the frequency of occurrence, the concept weight, the structural weight, and the corpus weight of the at least one concept.

11. (original): A method according to Claim 10, further comprising:

compressing the score through logarithmic compression.

12. (original): A method according to Claim 9, further comprising:

calculating the concept weight as a function of a number of terms comprising the at least one concept.

13. (original): A method according to Claim 9, further comprising:

calculating the structural weight as a function of a location of the at least one concept within the at least one such document.

14. (original): A method according to Claim 9, further comprising:

calculating the corpus weight as a function of a reference count of the at least one concept over the plurality of documents.

15. (original): A method according to Claim 9, further comprising:

forming the score assigned to the at least one concept to a normalized score vector for each such document;

determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector; and

applying the similarity to the best fit criterion.

16. (original): A method according to Claim 9, further comprising:

evaluating a set of candidate seed documents selected from the plurality of documents;

identifying a set of seed documents by applying the score for the at least one concept to a best fit criterion for each such candidate seed document; and

basing the best fit criterion on the score of each such seed document.

17. (currently amended): A computer-readable storage medium holding code for ~~performing the method of Claim 9.~~ grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:

code for determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is based on at least one of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight;

code for forming logically-grouped clusters of the documents by evaluating the score for the at least one concept of each document for a best fit to the clusters and assigning each document to the cluster with the best fit;

code for dynamically determining a threshold for each cluster based on similarities between the documents grouped into the cluster and a center of the cluster; and

code for reassigning those documents having similarities outside the threshold.

18. (currently amended): A system for providing efficient document scoring of concepts within [[a]] and clustering of documents in an electronically-stored document set, comprising:

a scoring module scoring a document in an electronically-stored document set, comprising:

a frequency module determining a frequency of occurrence of at least one concept within a ~~document retrieved from the document set; and~~ document;

a concept weight module analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document;

a structural weight module analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

a corpus weight module analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

a scoring evaluation module evaluating a score to be associated with the at least one concept as a function of the frequency, concept weight, structural weight, and corpus ~~weight.~~ weight; and

a clustering module grouping the documents by score into a plurality of clusters, comprising:

a cluster seed module identifying candidate seed documents, which are each assigned as a seed document into a cluster with a center most similar to the seed document, and assigning each non-seed document to the cluster with the best fit; and

a threshold module dynamically determining a threshold for each cluster based on similarities between the documents in each cluster and the cluster center, and reassigning the documents with similarities outside the threshold.

19. (currently amended): A system according to Claim 18, further comprising:

the scoring module evaluating the score ~~substantially~~ in accordance with the formula:

$$S_i = \sum_{i=n}^{i} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$ comprises the corpus weight for occurrence j of concept i.

20. (currently amended): A system according to Claim 19, further comprising:

the concept weight module evaluating the concept weight ~~substantially~~ in accordance with the formula:

$$cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

where cw_{ij} comprises the concept weight and t_{ij} comprises a number of terms for occurrence j of each such concept i.

21. (currently amended): A system according to Claim 19, further comprising:

the structural weight module evaluating the structural weight ~~substantially~~ in accordance with the formula:

$$sw_{ij} = \begin{cases} 1.0, & if(j \approx SUBJECT) \\ 0.8, & if(j \approx HEADING) \\ 0.7, & if(j \approx SUMMARY) \\ 0.5 & if(j \approx BODY) \\ 0.1 & if(j \approx SIGNATURE) \end{cases}$$

where sw_{ij} comprises the structural weight for occurrence j of each such concept i.

22. (currently amended): A system according to Claim 19, further comprising:

the corpus weight module evaluating the corpus weight ~~substantially~~ in accordance with the formula:

$$rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T \cdot}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \le M \end{cases}$$

where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for occurrence j of each such concept i, T comprises a total number of reference counts of documents in the document set, and M comprises a maximum reference count of documents in the document set.

23. (currently amended): A system according to Claim 19, further comprising:

a compression module compressing the score ~~substantially~~ in accordance with the formula:

$$S_i' = \log(S_i + 1)$$

where S_i' comprises the compressed score for each such concept i.

24. (original): A system according to Claim 18, further comprising:

a global stop concept vector cache maintaining concepts and terms; and

a filtering module filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache.

25. (original): A system according to Claim 18, further comprising:

a parsing module identifying terms within at least one document in the document set, and combining the identified terms into one or more of the concepts.

26. (original): A system according to Claim 25, further comprising:

the parsing module structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering.

27. (original): A system according to Claim 25, wherein at least one of nouns, proper nouns and adjectives are included as terms.

28. (original): A system according to Claim 18, further comprising:
a plurality of candidate seed documents;
a similarity module determining a similarity between each pair of a candidate seed document and a cluster center;
a clustering module designating each such candidate seed document separated from substantially all cluster centers with such similarity being sufficiently distinct as a seed document, and grouping each such candidate seed document not being sufficiently distinct into a cluster with a nearest cluster center.

29. (original): A system according to Claim 28, further comprising:
a plurality of non-seed documents;
the similarity module determining the similarity between each non-seed document and each cluster center; and
the clustering module grouping each such non-seed document into a cluster having a best fit, subject to a minimum fit criterion.

30. (original): A system according to Claim 29, further comprising:
a normalized score vector for each document comprising the score associated with the at least one concept for each such concept occurring within the document; and
the similarity module determining the similarity as a function of the normalized score vector associated with the at least one concept for each such document.

31. (currently amended): A system according to Claim 30, further comprising:
the similarity module calculating the similarity ~~substantially~~ in accordance with the formula:

$$\cos\sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A||\vec{S}_B|}$$

where $\cos\sigma_{AB}$ comprises a similarity between a document A and a document B, $\vec{S}_A$ comprises a score vector for document A, and $\vec{S}_B$ comprises a score vector for document B.

Claims 32-34 (canceled).

35. (currently amended): A method for providing efficient document scoring of concepts within [[a]] and clustering of documents in an electronically-stored document set, comprising:

scoring a document in an electronically-stored document set, comprising:

determining a frequency of occurrence of at least one concept within a ~~document retrieved from the document set; and~~ document;

analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document;

analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

evaluating a score to be associated with the at least one concept as a function of the frequency, concept weight, structural weight, and corpus ~~weight.~~ weight; and

grouping the documents by score into a plurality of clusters, comprising:

identifying candidate seed documents, which are each assigned as a seed document into a cluster with a center most similar to the seed document;

assigning each non-seed document to the cluster with the best fit;

dynamically determining a threshold for each cluster based on similarities between the documents in each cluster and the cluster center; and

reassigning the documents with similarities outside the threshold.

36. (currently amended): A method according to Claim 35, further comprising:

evaluating the score ~~substantially~~ in accordance with the formula:

$$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$ comprises the corpus weight for occurrence j of concept i.

37. (currently amended): A method according to Claim 36, further comprising:

evaluating the concept weight ~~substantially~~ in accordance with the formula:

$$cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

where cw_{ij} comprises the concept weight and t_{ij} comprises a number of terms for occurrence j of each such concept i.

38. (currently amended): A method according to Claim 36, further comprising:

evaluating the structural weight ~~substantially~~ in accordance with the formula:

$$sw_{ij} = \begin{cases} 1.0, & if(j \approx SUBJECT) \\ 0.8, & if(j \approx HEADING) \\ 0.7, & if(j \approx SUMMARY) \\ 0.5 & if(j \approx BODY) \\ 0.1 & if(j \approx SIGNATURE) \end{cases}$$

where sw_{ij} comprises the structural weight for occurrence j of each such concept i.

39. (currently amended): A method according to Claim 36, further comprising:

evaluating the corpus weight ~~substantially~~ in accordance with the formula:

$$rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for occurrence j of each such concept i, T comprises a total number of reference counts of documents in the document set, and M comprises a maximum reference count of documents in the document set.

40. (currently amended): A method according to Claim 36, further comprising:

compressing the score ~~substantially~~ in accordance with the formula:

$$S_i' = \log(S_i + 1)$$

where S_i' comprises the compressed score for each such concept i.

41. (original): A method according to Claim 35, further comprising:

maintaining concepts and terms in a global stop concept vector cache; and

filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache.

42. (original): A method according to Claim 35, further comprising:

identifying terms within at least one document in the document set; and

combining the identified terms into one or more of the concepts.

43. (original): A method according to Claim 42, further comprising:

structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering.

44. (original): A method according to Claim 42, further comprising:
including as terms at least one of nouns, proper nouns and adjectives.

Claim 45 (canceled).

46. (currently amended): A method according to Claim [[45,]] 35, further comprising:
identifying a plurality of non-seed documents;
determining the similarity between each non-seed document and each cluster center; and
grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit criterion.

47. (original): A method according to Claim 46, further comprising:
forming a normalized score vector for each document comprising the score associated with the at least one concept for each such concept occurring within the document; and
determining the similarity as a function of the normalized score vector associated with the at least one concept for each such document.

48. (currently amended): A method according to Claim 47, further comprising:
calculating the similarity ~~substantially~~ in accordance with the formula:

$$\cos\sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A||\vec{S}_B|}$$

where $\cos\sigma_{AB}$ comprises a similarity between a document A and a document B, $\vec{S}_A$ comprises a score vector for document A, and $\vec{S}_B$ comprises a score vector for document B.

Claims 49-51 (canceled).

52. (currently amended): A computer-readable storage medium holding code for ~~performing the method of Claim 35.~~ providing efficient document scoring of concepts within [[a]] and clustering of documents in an electronically-stored document set, comprising:

code for scoring a document in an electronically-stored document set, comprising:

code for determining a frequency of occurrence of at least one concept within a document;

code for analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document;

code for analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

code for analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

code for evaluating a score to be associated with the at least one concept as a function of the frequency, concept weight, structural weight, and corpus weight; and

code for grouping the documents by score into a plurality of clusters, comprising:

code for identifying candidate seed documents, which are each assigned as a seed document into a cluster with a center most similar to the seed document;

code for assigning each non-seed document to the cluster with the best fit;

code for dynamically determining a threshold for each cluster based on similarities between the documents in each cluster and the cluster center; and

code for reassigning the documents with similarities outside the threshold.

53. (currently amended): An apparatus for providing efficient document scoring of concepts within [[a]] and clustering of documents in an electronically-stored document set, comprising:

means for scoring a document in an electronically-stored document set, comprising:

means for determining a frequency of occurrence of at least one concept within a ~~document retrieved from the document set; and~~ document;

means for analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document;

means for analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

means for analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

means for evaluating a score to be associated with the at least one concept as a function of the frequency, concept weight, structural weight, and corpus ~~weight.~~ weight; and

means for grouping the documents by score into a plurality of clusters, comprising:

means for identifying candidate seed documents, which are each assigned as a seed document into a cluster with a center most similar to the seed document;

means for assigning each non-seed document to the cluster with the best fit;

means for dynamically determining a threshold for each cluster based on similarities between the documents in each cluster and the cluster center; and

means for reassigning the documents with similarities outside the threshold.